Analysis of two large functionally uncharacterized regions in the Methanopyrus kandleri AV19 genome

Jensen, Lars Juhl; Skovgaard, Marie; Sicheritz-Pontén, Thomas; Jørgensen, Merete Kjær; Lundegaard, Claus; Pedersen, Corinna Cavan; Petersen, Nanna; Ussery, David

Published in:
BMC Genomics

Link to article, DOI:
10.1186/1471-2164-4-12

Publication date:
2003

Document Version
Publisher's PDF, also known as Version of record

BMC Genomics

BioMed Central

Research article

Open Access

Analysis of two large functionally uncharacterized regions in the *Methanopyrus kandleri* AV19 genome

Lars Juhl Jensen, Marie Skovgaard, Thomas Sicheritz-Pontén, Merete Kjær Jørgensen, Christiane Lundegaard, Corinna Cavan Pedersen, Nanna Petersen and David Ussery*

Address: Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, The Technical University of Denmark, DK-2800 Lyngby, Denmark

Email: Lars Juhl Jensen - ljj@cbs.dtu.dk; Marie Skovgaard - marie@cbs.dtu.dk; Thomas Sicheritz-Pontén - thomas@cbs.dtu.dk; Merete Kjær Jørgensen - s011352@student.dtu.dk; Christiane Lundegaard - Christiane_lundegaard@mail.dk; Corinna Cavan Pedersen - s011395@student.dtu.dk; Nanna Petersen - s011604@student.dtu.dk; David Ussery* - Dave@cbs.dtu.dk

* Corresponding author

Published: 2 April 2003

BMC Genomics 2003, **4**:12

Received: 20 December 2002
Accepted: 2 April 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/12

Abstract

Background: For most sequenced prokaryotic genomes, about a third of the protein coding genes annotated are "orphan proteins", that is, they lack homology to known proteins. These hypothetical genes are typically short and randomly scattered throughout the genome. This trend is seen for most of the bacterial and archaeal genomes published to date.

Results: In contrast we have found that a large fraction of the genes coding for such orphan proteins in the *Methanopyrus kandleri* AV19 genome occur within two large regions. These genes have no known homologs except from other *M. kandleri* genes. However, analysis of their lengths, codon usage, and Ribosomal Binding Site (RBS) sequences shows that they are most likely true protein coding genes and not random open reading frames.

Conclusions: Although these regions can be considered as candidates for massive lateral gene transfer, our bioinformatics analysis suggests that this is not the case. We predict many of the organism specific proteins to be transmembrane and belong to protein families that are non-randomly distributed between the regions. Consistent with this, we suggest that the two regions are most likely unrelated, and that they may be integrated plasmids.

Background

Typically, for a newly sequenced genome the number of unique genes is mentioned. This is usually claimed to be about one third of the annotated genes. However, this number is highly questionable as random open reading frames (ORFs) are often assigned as protein coding genes. Based on an analysis of protein length distributions, we have estimated the true number of genes in each of the completely sequenced prokaryotic genomes [1]. Out of the estimated number genes in microbial genomes, *M. kandleri* contains the largest fraction of genes for which function cannot automatically be assigned based on sequence similarity (see supporting information at end of manuscript).

We have discovered that a significant fraction of these genes are located within two large regions of the chromosome (see Figure 1). This is in constrast to what is

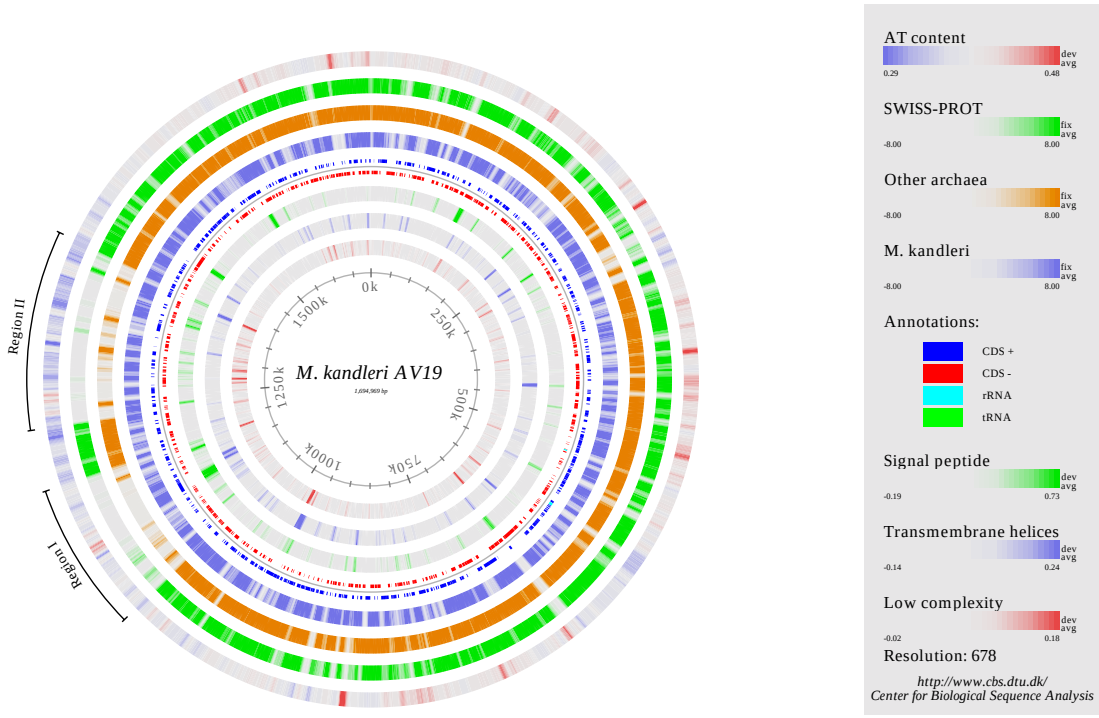

Figure 1
Atlas of the entire *M. kandleri* genome. Properties are shown as colored concentric circles representing the chromosome. These have all been smoothed by calculating 5,000 bp running averages.

observed in other prokaryotes where genes of unknown function are scattered throughout the genome. This organization of genes into large clusters would suggest that the proteins encoded by these genes are likely to represent novel protein complexes or biochemical pathways [2].

Pre-genome analysis of 16S ribosomal RNA had suggested *M. kandleri* to be placed close to the root of the Euryarchaeal tree. However, phylogenetic trees based on gene content, gene order, and ribosomal proteins places it toghether with the other archaeal methanogens [3,4]. It thus comes as a surprise that its genome appears to contain large numbers of genes not present in the genomes of any of the other sequenced archaeal methanogens. Furthermore, the *M. kandleri* genome has been claimed to contain very few genes acquired through lateral gene transfer [3]. It should however be noted that this claim was entirely based on an analysis of proteins with BLAST matches to sequences from other organisms. It is thus not possible to exclude that the two regions of unknown function could been transferred from other species that have not been characterized so far.

Results and discussion

To get an overview of the *M. kandleri* genome, we created circular visualizations known as genome atlases [5–7]. Figure 1 shows a customized atlas which summarizes the most interesting positional features of the *M. kandleri* genome: AT-content, predicted protein properties, and protein sequence similarity.

BLASTP searches of all predicted protein sequences in the *M. kandleri* genome were performed against a number of different databases. Comparison with SWISS-PROT was used for identifying proteins homologous with possible known function while a database of predicted archeal proteins was used for detecting conserved archeal proteins possibly missing in SWISS-PROT. The results of these two searches, as well as a search within the *M. kandleri* proteome, are shown in Figure 1.

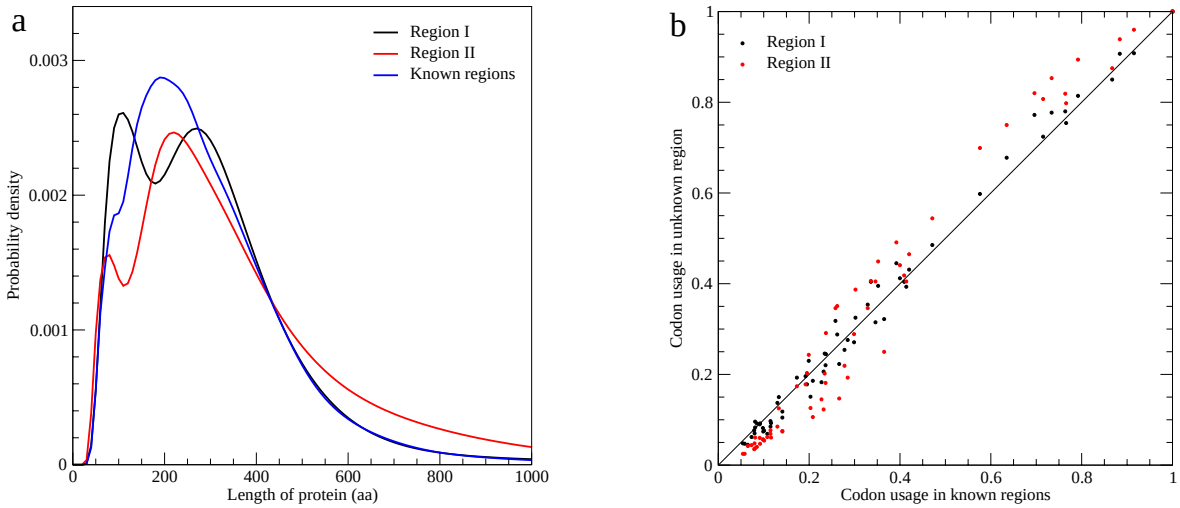

Figure 2
Comparison of protein length and codon usage for the unknown and known regions. (a) The length distributions of annotated proteins is visualized as Gaussian kernel density estimates. Based on the distributions we see no reason to suspect that the protein sequences from the two unknown regions are the result of random ORFs. (b) No differences in the relative usage of alternative codons for amino acids are observed between annotated CDSs from either of the two unknown regions and CDSs annotated in the known regions. This strongly indicates that the majority of the annotated novel genes in the unknown regions are true protein coding genes.

Figure 1 reveals two large chromosomal regions containing mostly protein coding genes without significant sequence similarity to genes from other organisms. The smaller of the two regions, *region I*, is located at 1,063 kbp (kilobase pairs) to 1,182 kbp and the slightly larger region (1,231 kbp–1,390 k) will be referred to as *region II*. To rule out that the genes within these regions are simply missing from the databases mentioned above, a TBLASTN search was performed against all sequences in GenBank. This resulted in no significant matches to DNA from other organisms than *M. handleri* itself.

From the atlas visualization it is clear that *region II* (but not *region I*) has a much lower AT-content than the genome average. The average AT-content is only 35.1% compared to the already low genomic average of 38.8%. An atypical local base composition is often used as supporting evidence for lateral gene transfer, although it should never be used alone [8].

Orphan proteins – not just ORFs

A very simple explanation for absence of homologous proteins in the two regions would be that the genes annotated are in fact not genes but merely random ORFs. As the majority of such sequences are characterized by being very short, this can be examined by studying the length distributions of annotated genes [1]. Figure 2(a) shows that the 210 genes annotated in the two regions follow distributions which are very similar to that of genes elsewhere in the M. kandleri genome. The only possible difference is that genes from region II tend to be slightly longer than other genes. The length distributions thus give us no reason to suspect that the annotated genes should be artifacts.

However, *region II* has an extremely low AT-content which can give rise to long random ORFs. It is therefore not possible, based on the length distributions alone, to completely rule out that the genes annotated in this region could be random ORFs. To further investigate this, the codon usage of genes in the unknown regions was compared to that of genes from known parts of the genome (Figure 2(b)). From the plot it is concluded that the codon preferences are identical in the unknown regions and other parts of the genome, despite large differences in amino acid composition (see below). Given this, we find it high-

ly unlikely that the annotated protein coding genes could be random ORFs. Furthermore, the similarity in codon usage also speak against the two regions having been acquired through lateral transfer of DNA. A more plausible explanation is that the two regions represent *M. kandleri* plasmids which have been integrated into the main chromosome.

Prediction of translation start

Discrimination between true protein coding genes and random ORFs is not the only problem in prokaryotic gene finding. Predicting the correct start codon is even more challenging as the ATG corresponding to the longest possible ORF is not necessarily the actual start codon used [9]. In addition to analysis of codon usage, modelling of the ribosomal binding site can also help determine the correct translation start site. This has already been taken into account by the gene finder used for annotating the *M. kandleri* genome [10].

The consensus sequence for the ribosome binding site (RBS) in an organism can generally be deduced from the 3' sequence of the 16S ribosomal RNA. In the *M. kandleri* AV19 genome only one 16S rRNA sequence is annotated, ending with the sequence CACCTC-3'. However, a sequence comparison with 16S sequences from the Ribosomal Database Project [11] revealed that the most closely related sequences all end on CACCTCC-3'. As the first nucleotide after the annotated 16S rRNA in *M. kandleri* genome is indeed a cytosine, we suggest that the annotated 16S rRNA is missing one nucleotide. The corresponding RBS consensus sequence would thus be GGAGGTG, the reverse complement of the rRNA 3' end.

To further examine the RBS sequence, regions of 30 bp immediately upstream of each annotated coding region were examined for overrepresented sequence patterns. Surprisingly, stretches of four to six guanines rather than the anticipated RBS consensus sequences turned out to be the most significantly overrepresented patterns. The anticipated RBS sequence, GGAGG, was found to be significantly overrepresented in the positive set at $P < 10^{-6}$ while the patterns GGGG, GGGGG, and GGGGGG were all significant at $P < 10^{-10}$ or better. Figure 3 shows the positioning of the patterns GGGGG and GGAGG relative to the annotated translation starts. While both patterns exhibit a clear preference for occuring at a distance of 10 to 15 bp from the start codon, this preference is strongest for the pattern GGGGG. No difference in the RBS preference is observed between the unknown and known regions, which is consistent with evidence against lateral gene transfer provided by codon usage analysis.

That GGGGG is the most overrepresented pattern just upstream of translation start would suggest that it is the preferred RBS sequence despite it not having perfect complementarity to the 16S rRNA 3' end. The mismatch between the 3'-end of the 16S rRNA and the most common RBS can easily be accommodated as it corresponds to the wobble base pairing between guanine and uracil, which is very common in RNA [12].

Protein families in **M. kandleri**

Although the vast majority of proteins from the two regions described earlier have no significant similarity to proteins from other organisms, similarities exist among *M. kandleri* proteins. Based on this, 45 *M. kandleri* specific protein families were defined by Slesarev et al [3]. Several of these protein families are non-randomly distributed between the two unknown regions and known regions: all five members of the MK-10 protein family are encoded by genes in *region II* while *region I* contains all five genes encoding MK-9 proteins. Other protein families with skewed occurences can be found in Table 1.

The pairwise similarities between members of the same *M. kandleri* specific protein family gives rise to extensive homology to other *M. kandleri* proteins as seen in Figures 1 and 4. Many of the protein families show a tendency to be encoded by clusters of genes occurring on the same strand, *e.g.* the MK-5, MK-9, MK-10, and MK-28 families labelled in Figure 4. This, combined with the fact that the homology between proteins from the *M. kandleri* families is strong enough to also be detectable at the DNA level, suggest that the families are likely to represent recent gene duplication events.

Properties of proteins from the unknown regions

The localization of different *M. kandleri* specific protein families in the two unknown regions (Table 1) together with the difference in AT-content (Figure 1) suggests that the two regions should be studied separately. To do so, linear zooms of the two regions were constructed (Figure 4).

Several striking subregions can be found within both of the unknown regions. *Region I* is dominated by proteins predicted to have N-terminal signal peptides, which strongly suggests them to be secreted (see Figure 4). However, this is largely due to two very large proteins from the MK-5 family located at 1,126 kbp to 1,136 kbp. In addition to this region, several other clusters of proteins predicted to have signal peptides are observed within both *region I* and *region II*.

In addition to clusters of secreted proteins, several groups of predicted transmembrane proteins are observed in both regions – in particular *region I*. Some degree of overlap is observed between proteins predicted to be transmembrane and those predicted to have signal pep-

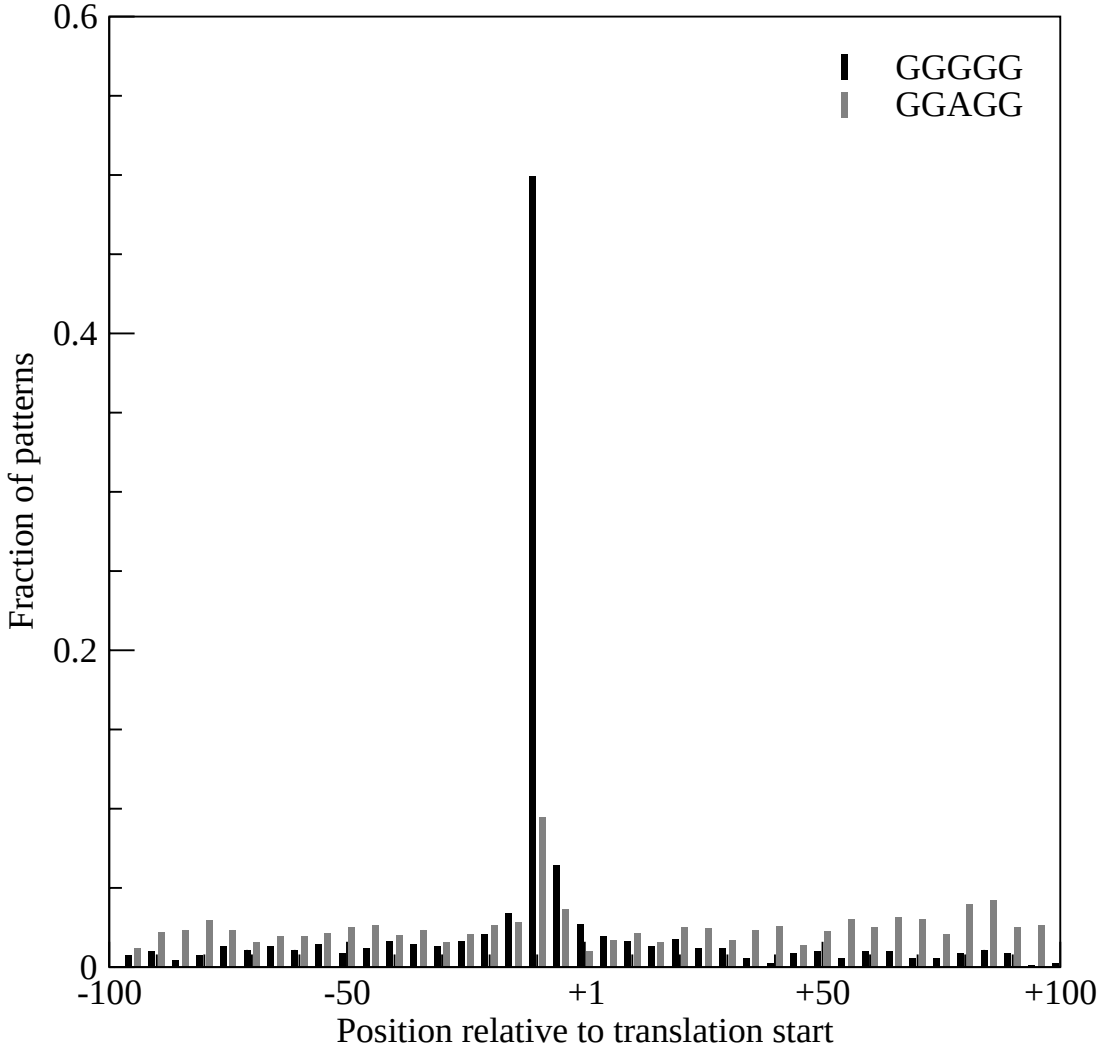

Figure 3
Position of the patterns GGGGG and GGAGG relative to translation start. Despite GGAGG being the reverse complement of the 3' end of *M. kandleri* 16S rRNA, the pattern GGGGG is found to have a much stronger preference for being located just upstream of translation start.

tides. The MK-9 protein family occurring exclusively within *region I* constitute one cluster of transmembrane proteins (1,138 k–1,149 k). Similarly, the five members of the MK-10 family in *region II* form a cluster of proteins predicted to be transmembrane (1,375 k–1,382 k). Considering the special membrane of *M. kandleri* which consists of a terpenoid lipid [13] and the extreme conditions under which it lives, the presence of special membrane proteins is hardly surprising.

Three of the MK-10 family proteins also contain low-complexity regions. In addition to these proteins, *region II* con-

Table 1: Distribution of *M. kandleri* specific protein families. Only the subset of the 45 protein families showing a preference for either region I or II is shown. The presence of specific protein families within each region suggests that the two regions serve different functions.

Family	Region I	Region II	Known
MK-9	5	-	-
MK-7	3	-	1
MK-6	3	-	4
MK-5	3	1	4
MK-17	2	-	-
MK-23	2	-	-
MK-26	2	-	-
MK-27	2	-	-
MK-8	2	-	1
MK-22	2	-	1
MK-1	3	6	11
MK-10	-	5	-
MK-2	-	4	1
MK-3	-	4	3
MK-11	-	3	-
MK-12	1	3	1
MK-14	-	3	2
MK-37	-	3	3
MK-31	-	2	-
MK-34	-	2	1
MK-28	-	2	2

tains an abundance of other low-complexity proteins that are, however, not predicted to be transmembrane. Low-complexity regions often form unstructured non-globular domains [14,15]. DNA binding proteins have previously been shown to contain more low-complexity regions than other proteins [16]. It is thus tempting to speculate that some of the low-complexity proteins found in *M. kandleri* could be involved in stabilizing DNA at extreme temperatures.

Amino acid biases

Low-complexity regions are defined as regions with strong bias towards one or more amino acid residues. It thus possible to further characterize regions of low-complexity by studying which residues are overrepresented. Such biased regions will often lead to an unusual amino acid composition of the protein as such.

We have compared the amino acid composition of proteins from each of the two unknown regions to that of proteins from other parts of the chromosome. The amino acid biases for the two regions are shown in Figure 5. The residue with the strongest positive bias is tryptophan, which is about 80% more frequent within the two regions compared to elsewhere. Other residues found to be overrepresented are serine, proline, and leucine. Additionally, arginine is more frequent in *region II* while threonine is overrepresented in *region I*. Atlas visualizations of amino acid composition revealed that these positive biases are mainly due to single highly biased proteins, rather than a general trend for the regions (data not shown).

Also, Figure 5 shows an underrepresentation of methionine in proteins from both regions, as well as a strong bias against isoleucine and lysine in *region II*. In contrast to the positive biases, the negative ones appear to be due to a relatively weak bias in the majority of the proteins. The residues found to be underrepresented in *region II* have in common that they are encoded by AT-rich codons. This is consistent with the lower AT-content of *region II* compared to the rest of the genome.

This leads to an interesting "hen and egg" problem: is it the low AT-content of *region II* that is the cause of the biases in amino acid composition or vice versa? The lack of a positive bias for amino acid residues encoded by GC-rich codons would suggest the latter. This hypothesis is further supported by an analysis of AT-content of intergenic regions within *region II* to that of intergenic regions from other parts of the chromosome, which revealed no significant difference according to a Kolmogorov-Smirnov test [17].

A link between the amino acid compositions of proteins and their thermostability has previously been suggested by several groups [18–20]. There is little agreement

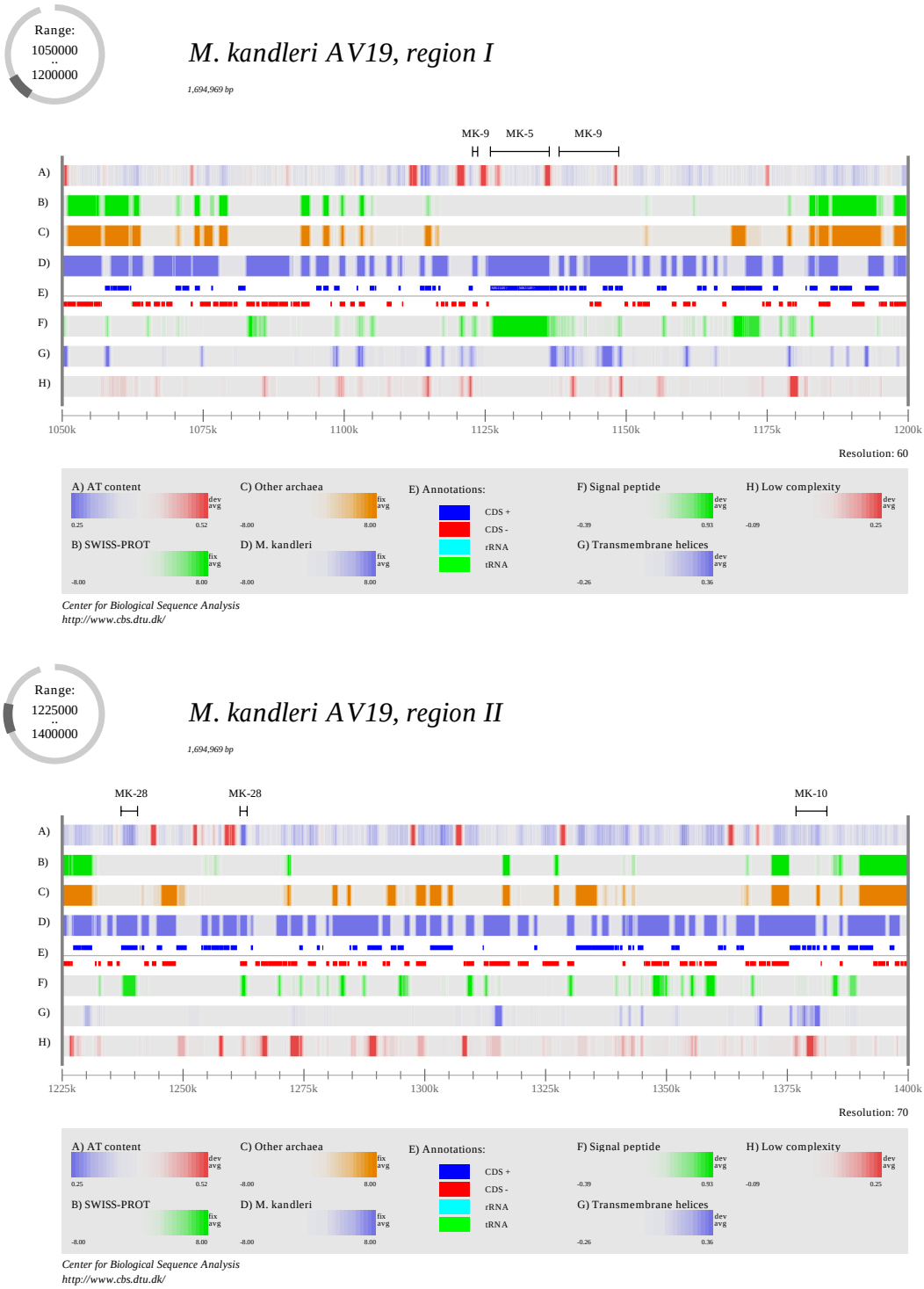

Figure 4
Atlases of the two unknown regions. The properties are visualized as in Figure 1 except that a smoothing window of 1,000 bp was applied to all parameters.

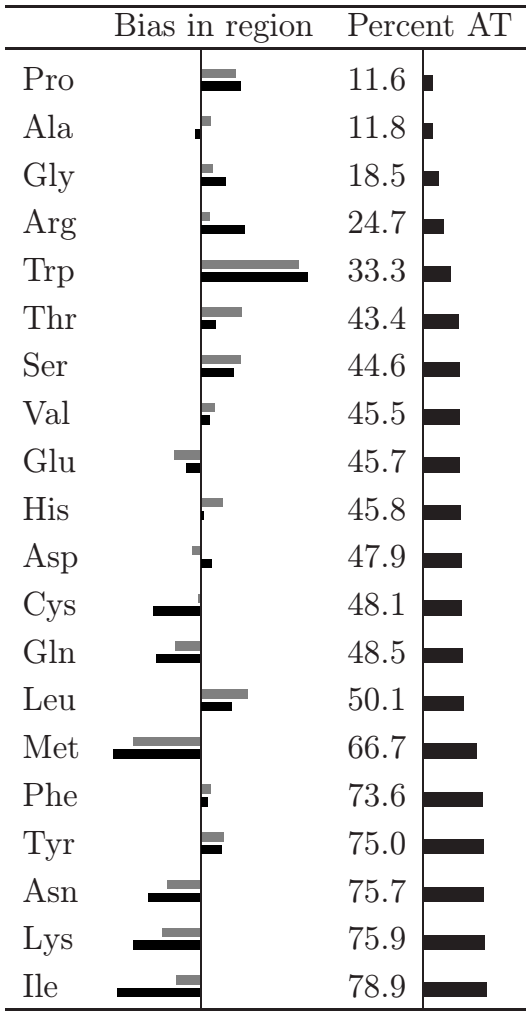

Figure 5
Amino acid compositional biases in the unknown regions. The bias of each amino acid is represented by a bar with length proportional to the log-ratio between its amino acid frequency within one of the unknown regions and its frequency within known regions. Regions I and II are represented by gray and black bars respectively. The amino acids have been sorted by their codon AT-content, which is both listed and visualized as bars.

between the compositional characteristics of thermostable proteins suggested, although asparagine and glutamine are generally found to be underrepresented while arginine is overrepresented. Although this is consistent with the amino acid biases observed in *region II*, the correlation is very weak as the strongest biases do not correlate with thermostability. We therefore see no reason to suspect proteins from either of the two unknown regions to have different thermostability than other *M. kandleri* proteins.

Conclusions

While claims of newly sequenced genomes containing large numbers of unique proteins are often encountered, the existence of these proteins is rarely backed up by anything but a gene finding method. While such spurious gene predictions are typically scattered throughout the genome, we have discovered two distinct, large regions in the genome of *M. kandleri* where only a small fraction of the annotated genes share significant similarity with known proteins from other organisms. Analysis of length distributions and codon usage strongly suggests the presence of a large number of unique protein coding genes within these regions and rejects the hypothesis that the regions could have been aquired through lateral gene transfer. Instead, we believe the most likely origin to be integration of plasmids. Extensive bioinformatics analysis of the proteins encoded by these regions suggests many of them to be transmembrane, but little else can be predicted about their functions. Additional experimental data is needed in order to learn more about these proteins.

Methods

The genome sequence and annotation of the *M. kandleri* AV19 genome was downloaded from GenBank [21] along with 13 other sequenced archaeal genomes (*A. pernix*, *P. aerophilum*, *S. solfataricus*, *S. tokodaii*, *T. acidophilum*, *T. volcanium*, *A. fulgidus*, *M. thermoautotrophicum*, *M. jannaschii*, *M. acetivorans*, *M. mazei*, *P. abyssi*, and *P. horikoshii*).

Sequence similarity searches

All sequence similarity searches were performed using gapped BLAST with low complexity filter enabled [22]. The conceptual translations of all annotated protein coding genes in the *M. kandleri* genome were searched against four different sequence databases: the *M. kandleri* proteome itself, SWISS-PROT [23], GenBank [21], and a set of all annotated protein coding genes from the 13 other sequenced archaeal genomes. In the BLAST results from the search against the *M. kandleri* proteome, the self match of each protein was discarded.

Prediction of protein properties

Protein properties were predicted from sequence using a wide array of prediction methods. Low-complexity regions in the sequences were identified using SEG, which is the program used by BLAST to mask such regions [24]. SignalP was used to predict signal peptides using the model trained on eukaryotic proteins [25]. Transmembrane helices were predicted using the TMHMM method [26]. These three protein features were the only that were

Table 2: Fraction of proteins assignable to cellular role categories. The estimated number of genes in each genome is compared to the number of genes for which a cellular role could be assigned using EUCLID.

Organism	No. genes estimated	No. genes assigned	%of estimate
M. kandleri	1,477	653	44
A. pernix	1,376	684	50
P. aerophilum	1,706	867	51
S. tokodaii	2,035	1,045	51
S. solfataricus	2,288	1,186	52
M. mazei	2,686	1,420	53
M. acetivorans	3,456	1,850	54
P. furiosus	1,683	911	54
P. horikoshii	1,448	786	54
Halobacterium sp.	1,573	895	57
A. fulgidus	1,818	1,074	58
M. jannaschii	1,350	781	58
P. abyssi	1,497	855	58
M. thermoautotrophicum	1,466	867	60
T. acidophilum	1,250	783	63
T. volcanium	1,243	792	64

found to correlate significantly with the unknown regions according to Kolmogorov-Smirnov tests [17].

The other features tested for correlations were: grand average hydropathity [27], instability index [28], predicted glycosylation sites [29], predicted phosphorylation sites [30], PEST regions [31], and secondary protein structure predicted by PSIPRED [32].

Atlas visualization

The atlas visualization is a circular representation of a microbial genome in which different DNA or protein properties are visualized as colored circles [5–7].

The matches found by the BLAST searches described above are visualized as three circles (Figure 1) representing each of the databases against which BLAST searches were performed. For every protein the negative logarithm of E-value of the most significant match was calculated (imposing a maximum score of 15 for highly significant matches). These values were mapped to the chromosomal location of the corresponding genes color coded so that regions with many significant matches are colored whereas regions with few matches are gray.

The predicted protein properties were plotted by making similar mappings of the predicted fractions of transmembrane and low-complexity residues in each protein. Signal peptide predictions were represented by their mean S-score [25]. All three sets of values were color coded using a scheme so that only regions containing unusually many secreted, transmembrane, and/or low-complexity proteins will be visible.

Finally, we include a circle showing the AT-content (which is closely correlated to many DNA structural properties [6]). A double sided color scheme was used which highlights regions of unusually high or low AT-content compared to the average for the genome.

Length distribution plots

Protein length distributions for proteins from different regions of the M. kandleri were plottes as density estimates. Rather than using simple histograms, Gaussian kernel density estimates were calculated for log-transformed protein lengths. The widths of the Gaussian kernels were estimated based on the number of data points and their spread [33]. The log-length density estimates were subsequently transformed back to yield ordinary length distributions.

Calculation of codon usage and AT-content

For genes residing in the each of the unknown regions and for genes from known regions, the codon usage was calculated by first counting the frequency of each of the 61 coding triplets. The triplets encoding each amino acid were then normalized to a sum of one to cancel effects due to the amino acid composition. We refer to these normalized frequencies as the codon usage. Applying the codon usage estimated from all annotated protein coding regions as weighting factors, the average codon AT-content for each amino acid was calculated as a weighted average of the AT-content of all codons encoding the amino acid in question.

RBS pattern search

The sequences located at positions -30 to -1 relative to translation start were extracted for all annotated protein coding genes. These sequence were used as positive examples in the subsequent search. The sequences for positions -60 to -31, again relative to translation start, were extracted for use as negative examples. All DNA words up to a length of 10 bp were tested for significant overrepresentation in the positive examples relative to the negative examples using a hypergeometric test as described by Jensen and Knudsen [34].

Authors' contributions

• Lars Juhl Jensen: Worked on the analysis of protein length distributions and codon usage. Discovered the alternative RBS consensus sequence. Wrote substancial parts of the manuscript.

• Marie Skovgaard: Was involved in length distribution and codon usage analysis. Wrote large parts of the manuscript.

• Thomas Sicheritz-Pontén: Initially discovered the two regions together with David Ussery. Participated in analysis of possible lateral gene transfer.

• Merete Kjær Jørgensen and Christiane Lundegaard: Generation and analysis of BLAST results and atlas visualizations and detailed analysis of *Region I*.

• Corinna Cavan Pedersen and Nanna Petersen: Analysis of amino acid composition of proteins from the two regions. Detailed examination of atlases for *Region II*.

• David Ussery: Initial discovery of the two regions. Coordination of the project and contributed in the final stages of manuscript preparation.

Supporting information

To examine the fraction of genes with known function in each of the sequenced prokaryotic genomes, the EUCLID method was used for automatically assigning cellular role categories to annotated proteins based on BLASTP matches to SWISS-PROT [22,23,35,36]. The number of proteins in each genome for which a category could be assigned was compared to our estimate of the number of protein coding genes in the genome [1]. Table 2 shows the results of this analysis, which reveals that *M. kandleri* is the genome for which function can be assigned to the smallest fraction of the estimated number of protein coding genes. This is largely due to the two regions of unknown function.

Acknowledgements

This work was supported by grants from the Danish National Research Foundation and the Danish Natural Science Research Council. Marie Skovgaard is funded by EU Cell Factory Project, Screen, QLK3-CT-2000-00649.

References

1. Skovgaard M, Jensen L, Brunak S, Ussery D and Krogh A **On the total number of genes and their length distribution in complete micro-bial geuomes** *Trends in Genetics* 2001, **17:**425-428
2. Rogozin I, Makarova K, Murvai J, Czabarka E, Wolf Y, Tatusov R, Szekely L and Koonin E **Connected gene neighborhoods in prokarytic genomes** *Nucl Acids Res* 2002, **30:**2212-2223
3. Slesarev A, Mezhevaya K, Makarova K, Polushin N, Shcherbinina O, Shakhova V, Belova G, Aravind L, Natale D and Rogozin I **The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens** *Proc Natl Acad Sci USA* 2002, **99:**4644-4649
4. Sebaihia M, Bentley S, Thomson N, Holden M and Parkhill J **Tales of the unexpected** *Trends in Microbiology* 2002, **10:**261-262
5. Jensen L, Friis C and Ussery D **Three views of microbial genomes** *Res Microbiol* 1999, **150:**773-777
6. Pedersen A, Jensen L, Stasrfeldt H, Brunak S and Ussery D **A DNA structural atlas of *E. coli*** *J Mol Biol* 2000, **299:**907-930
7. Skovgaard M, Jensen L, Friis C, Stærfeldt HH, Worning P, Brunak S and Ussery D **The atlas visualisation of genome-wide information** *In, Methods in Microbiology (Edited by: Wren B, Dorrell N) Academic Press, London, UK* 2002, **33:**49-63
8. Ragan M **On surrogate methods for detecting lateral gene transfer** *FEMS Microbiol Lett* 2001, **201:**187-191
9. Hannenhalli S, Hayes W, Hatzigeorgiou A and Fickett J **Bacterial start prediction** *Nucl Acids Res* 1999, **27:**3577-3582
10. Besemer J, Lomsadze A and Borodovsky M **GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions** *Nucl Acids Res* 2001, **29:**2607-2618
11. Maidak B, Cole J, Lilburn T, Parker C Jr, Saxman P, Farris R, Garrity G, Olsen G, Schmidt T and Tiedje J **The RDP-II (Ribosomal Database Project)** *Nucl Acids Res* 2001, **29:**173-174
12. Gautheret D, Konings D and Gutell R **G: U base pairing motifs in ribosomal RNA** *RNA* 1995, **1:**807-814
13. Hafenbradl D, Keller M, Thiericke R and Stetter K **A novel unsaturated archaeal ether core lipid from the hyperthermophile *Methanopyrus kandleri*** *Syst Appi Microbiol* 1993, **16:**165-169
14. Wright P and Dyson H **Intrinsically unstructured proteins: Reassessing the protein structure – function paradigm** *J Mol Biol* 1999, **293:**321-331
15. Dunker A and Obradovic Z **The protein trinity – linking function and disorder** *Nature Biotechnology* 2001, **19:**805-806
16. Wise M **Ojpy: a software tool or low complexity proteins and protein domains** *Bioinformatics* 2001, **17:**S288-S295
17. Young I **Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources** *J Histochem Cytochem* 1977, **25:**935-941
18. Haney P, Badger J, Buldak G, Reich C, Woese C and Olsen G **Thermal adaption analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species** *Proc Nail Acad Sci USA* 1999, **96:**3578-3583
19. Kreil D and Ouzounis C **Identification of thermophilic species by the amino acid composition deduced from their genomes** *Nucl Acids Res* 2001, **29:**1608-1615
20. Kumar S and Nussinov R **How do thermophilic proteins deal with heat?** *Cell Mol Life Sci* 2001, **58:**1216-1233
21. Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Rapp B and Wheeler D **GenBank** *Nucl Acids Res* 2002, **30:**17-20
22. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W and Lipman D **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs** *Nucl Acids Res* 1997, **25:**3389-3402
23. Bairoch A and Apweiler R **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000** *Nucl Acids Res* 2000, **28:**45-48
24. Wootton J and Federhen S **Statistics of local complexity in amino-acid-sequences and sequence data bases** *Comput Chem* 1993, **17:**149-163

25. Nielsen H, Brunak S and von Heijne G **Machine learning approaches for the prediction of signal peptides and other protein sorting signals** *Protein Eng* 1999, **12:**3-9
26. Krogh A, Larsson B, von Heijne G and Sonnhammer E **Predicting transmembrane protein topology with a hidden markov model: application to complete genomes** *J Mol Biol* 2001, **305:**567-580
27. Kyte J and Doolittle R **A simple method for displaying** *J Mol Biol* 1982, **157:**105-132
28. Guruprasad K, Reddy B and Pandit M **Correlation between stability of a protein and its di-peptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence** *Protein Eng* 1990, **4:**155-161
29. Hansen J, Lund O, Tolstrup N, Gooley A, Williams K and Brunak S **tOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility** *Glycoconj J* 1998, **15:**115-130
30. Blom N, Gammeltoft S and Brunak S **Sequence and structure-based prediction of eukaryotic protein phosphorylation sites** *J Mol Biol* 1999, **294:**1351-1362
31. Rechsteiner M and Rogers S **PEST sequences and regulation by pro-teolysis** *Trends Biochem Sci* 1996, **21:**267-271
32. Jones D **Protein secondary structure prediction based on position-specific scoring matrices** *J Mol Biol* 1999, **292:**195-202
33. Silverman B **Density Estimation for Statistics and Data Analysis** *Chapman & Hall, London* 1986, Chap 3
34. Jensen L and Knudsen S **Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation** *Bioinformatics* 2000, **16:**326-333
35. Tamames J, Ouzounis C, Casari G, Sander C and Valencia A **EUCLID: automatic classification of proteins in functional classes by their database annotations** *Bioinformatics* 1998, **14:**542-543
36. Andrade M, Brown N, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A and Ouzounis C **Automated genome sequence analysis and annotation** *Bioinformatics* 1999, **15:**391-412